OPTIMUM RULES FOR CLASSIFICATION INTO TWO MULTIVARIATE NORMAL POPULATIONS WITH THE SAME COVARIANCE MATRIX

by

Somesh Das Gupta *

University of Minnesota

Technical Report No. 359

October 1979

* Supported by a grant from the Mathematics Division, U.S. Army Research Office, Durham, N.C. Grant No. DAAG-29-0038.

1. Introduction.

Let w denote an experimental unit drawn randomly from a population π. The classification problem in its standard form is to devise rules so as to identify π with one of the two given "distinct" populations π_1 and π_2. A set of p real-valued measurements $\underset{\sim}{X}: p \times 1$ is observed on w and it is believed that the distributions of $\underset{\sim}{X}$ in those two populations are different. In this paper we shall assume that $\underset{\sim}{X} \sim N_p(\underset{\sim}{\mu}, \Sigma)$.

Let $\underset{\sim}{\mu}_i$ denote the mean of $\underset{\sim}{X}$ in the population π_i $(i = 1,2)$, where $\underset{\sim}{\mu}_1 \neq \underset{\sim}{\mu}_2$. The classification problem is to find "good" rules for deciding whether $\mu = \mu_1$ or $\mu = \mu_2$. When all the parameters μ_1, μ_2 and Σ are known Wald's decision theory [17] may be used to derive the miminal complete class of decision rules for zero-one loss function. It is given by the following, except for sets of measure zero [2]: The rule ϕ^k decides $\mu = \mu_1$ iff

$$(1.1) \qquad (x - \mu_1)' \Sigma^{-1} (x - \mu_1) - (x - \mu_2)' \Sigma^{-1} (x - \mu_2) \leq k$$

It can be proved [2] that the rule ϕ^o is the only admissible minimax rule.

However, in practice all the parameters are not known, and in order to differentiate the two populations random (training) samples from both the populations are obtained. It may be remarked that if either of μ_1 and μ_2 is known it is not necessary to draw samples from both the populations.

Let θ stand for $(\mu, \mu_1, \mu_2, \Sigma)$, and

(1.2) $\Theta_1 = \{\theta: \mu = \mu_1, (\mu_1, \mu_2, \Sigma) \in \Omega\}$,

(1.3) $\Theta_2 = \{\theta: \mu = \mu_2, (\mu_1, \mu_2, \Sigma) \in \Omega\}$,

where Ω is a known set in the space of μ_1, μ_2 and Σ. It may be noted that in order to control (arbitrarily) both probabilities of incorrect classification certain conditions must be imposed on Ω and sequential sampling schemes may have to be used [5]. However, in standard practice Ω is taken to be the set

(1.4) $\Omega = \{(\mu_1, \mu_2, \Sigma): \mu_1, \mu_2, \in R^p, \mu_1 \neq \mu_2, \Sigma$ is positive-definite$\}$.

Following Fisher [7] a set of heuristic rules (called plug-in rules) may be devised by first choosing some good estimates of the unknown parameters and replacing the unknown paramaters in ϕ^k by their respective estimates. We shall call such a rule ϕ_p^k when the standard estimates are used.

Let $X_{i1}, \ldots, X_{in_i}$ denote the X-oberservations of the training sample from π_i $(i = 1, 2)$. Define (assume $n_1 + n_2 - 2 > 0$)

(1.5) $$\bar{X}_i = \sum_{j=1}^{n_1} X_{ij}/n_i \quad (i = 1,2),$$

$$S = [\sum_{j=1}^{n_1} (X_{ij} - \bar{X}_1)(X_{ij} - \bar{X}_1)' + \sum_{j=1}^{n_2} (X_{2j} - \bar{X}_2)(X_{ij} - \bar{X}_2)']/(n_1+n_2-2)$$

When all the parameters are unknown, Fisher's plug-in rules are given by the following: The rule ϕ_p^k decides $\mu = \mu_1$ iff

$$(1.6) \qquad (X - \bar{X}_1)' S^{-1} (X - \bar{X}_1) - (X - \bar{X}_2)' S^{-1} (X - \bar{X}_2) \leq k$$

Using the likelihood-ratio principle Anderson [1] proposed the following rules when (μ_1, μ_2, Σ) lies in Ω given by (1.4): The rule ψ_L^λ decides $\mu = \mu_1$ iff

$$(1.7) \qquad (1 + 1/n_1)^{-1} (X - \bar{X}_1)' S^{*-1} (X - \bar{X}_1) - \lambda (1 + 1/n_2)^{-1}$$

$$(X - \bar{X}_2)' S^{*-1} (X - X_2) \leq \lambda - 1,$$

where $S^* = mS$, $m = n_1 + n_2 - 2$. $(\lambda > 0)$ Note that $\psi_L^1 = \phi_p^0$ when $n_1 = n_2$. The likelihood-ratio rules turn out to be the following when Σ is known: The rule ϕ_L^k decides $\mu = \mu_i$ iff

$$(1.8) \qquad (1 + 1/n_1)^{-1} (X - \bar{X}_1)' \Sigma^{-1} (X - \bar{X}_1)$$

$$- (1 + 1/n_2)^{-1} (X - \bar{X}_2)' \Sigma^{-1} (X - \bar{X}_2) \leq k.$$

One may also derive some "good" constructive rules from various optimality criteria. In this paper we shall obtain some good rules from Wald's decision-theoretic viewpoint, and also from asymmetrical Neyman-Pearson approach. We shall also study the above two classes of heuristic rules from some optimality criteria.

2. The Univariate Case

2-1. $p = 1$, σ^2 is known. Without any loss of generality we shall assume that $\sigma^2 = 1$. Let $\varphi = (\varphi_1, \varphi_2)$ stand for a decision rule, where φ_i

is the probability of deciding $\mu = \mu_i$ given the observations. We shall consider only the rules based on sufficient statistics X, $\bar{X}_1$ and $\bar{X}_2$.

First we shall make an orthogonal transformation as follows: Define

$$(2.1) \qquad U_1 = k_1[(1+1/n_1)^{-\frac{1}{2}}(X-\bar{X}_1) + (1+1/n_2)^{-\frac{1}{2}}(X-\bar{X}_2)],$$

$$(2.2) \qquad U_2 = k_2[(1+1/n_1)^{-\frac{1}{2}}(X-\bar{X}_1) - (1+1/n_2)^{-\frac{1}{2}}(X-\bar{X}_2)],$$

$$(2.3) \qquad U_3 = k_3[X + n_1\bar{X}_1 + n_2\bar{X}_2],$$

where k_i's are chosen so that $\text{var}(U_i) = 1$; $i = 1,2,3$. Note that U_i's are independently distributed. Let $E(U_i) = \nu_i$. Then $U_i \sim N(\nu_i,1)$.

In terms of (ν_1, ν_2, ν_3) the sets Θ_1 and Θ_2 as defined in (1.2)-(1.4), are transformed as follows:

$$(2.4) \qquad \Omega_1 = \{(\nu_1,\nu_2,\nu_3): \nu_1 = \nu, \nu_2 = -c\nu, \nu \neq 0, \nu_3 \in R\}$$

$$(2.5) \qquad \Omega_2 = \{(\nu_1,\nu_2,\nu_3): \nu_1 = \nu, \nu_2 = c\nu, \nu \neq 0, \nu_3 \in R\} ,$$

where $c = k_2/k_1 > 0$. (k_i's are chosen to be positive.) Note that $c > 1$.

2.1.1 Bayes Rules and Minimax Rules

It is easy to see that by taking a suitable prior distribution of ν_3 independently of ν_1 and ν_2 we can get Bayes rules free from U_3. Hence we shall only consider prior distributions of (ν_1,ν_2) and drop U_3 from the argument of φ. Let

$$(2.6) \qquad \Omega_i^* = \{(\nu_1,\nu_2): \nu_1 = \nu, \nu_2 = (-1)^i c\nu, \nu \neq 0\} .$$

Consider a prior distribution $\xi(\beta,\gamma,\nu_0)$ which assigns probabilities $\beta\gamma$, $(1-\beta)(1-\gamma)$, $\beta(1-\gamma)$, $\gamma(1-\beta)$ to the parameter points $(\nu_0, c\nu_0)$, $(-\nu_0,-c\nu_0)$, $(\nu_0, -c\nu_0)$, $(-\nu_0, c\nu_0)$, respectively, where $0 \leq \beta \leq 1$, $0 \leq \gamma \leq 1$, $\nu_0 > 0$. It can be seen that the unique (a.e.) Bayes rule (for zero-one loss function) against the above prior distribution is given by the following: Decide $(\nu_1, \nu_2, \nu_3) \in \Omega_1$ iff

(2.7) $(U_1-c_1)(U_2-c_2) \leq 0$,

where c_1 and c_2 are functions of ν_0, β, γ and c. Conversely, given c_1 and c_2 it is possible to choose β,γ and ν_0 appropriately. Another class of Bayes rules may be obtained from the following prior distributions: The probability that $(\nu_1,\nu_2) \in \Omega_i^*$ is ξ_i, and given that $\nu_1 = \nu$, $\nu_2 = (-1)^i c\nu$ the distribution of ν is $N(0, \tau^2)$. The unique (a.e.) Bayes rule against the above prior distribution decides $(\nu_1,\nu_2,\nu_3) \in \Omega_1$ iff

(2.8) $U_1U_2 \leq k$,

where k is a function of ξ_1, ξ_2 and c. Different types of Bayes rules are given by DasGupta and Bhattacharya [3].

Now consider the rule which decides $(\nu_1,\nu_2,\nu_3) \in \Omega_1$ iff

(2.9) $U_1U_2 \leq 0$.

Note that (2.9) is equivalent to

(2.10) $(1+1/n_1)^{-1}(X-\bar{X}_1)^2 \leq (1+1/n_2)^{-1}(X-\bar{X}_2)^2$.

Thus the above rule is the same as φ_L^0, defined in (1.8). The rule φ_L^0 is the unique Bayes rule against the prior $\xi(\frac{1}{2},\frac{1}{2},\nu_0)$ for any $\nu_0 > 0$. Moreover, the risk of the rule φ_L^0 is constant over the four-point set $(\nu_0,c\nu_0)$, $(-\nu_0,-c\nu_0)$, $(-\nu_0,c\nu_0)$, $(\nu_0,-c\nu_0)$. Hence φ_L^0 is an admissible minimax rule, and moreover the supremum of the risk of φ_L^0 is equal to $\frac{1}{2}$.

However, φ_L^0 is not the unique minimax rule (leaving aside the trivial rule $\varphi_1 \equiv \varphi_2 \equiv \frac{1}{2}$). To see this, transform (U_1,U_2) to (V_1,V_2) by an orthogonal transformation L such that (EV_1,EV_2) is proportional to $(1, -d_1)$ and $(1,d_2)$ for $(\nu_1,\nu_2) \in \Omega_1^*$ and $(\nu_1,\nu_2) \in \Omega_2^*$, respectively, and $d_1 > 0$, $d_2 > 0$. Let ψ be the rule which decides $(\nu_1,\nu_2) \in \Omega_1^*$ iff $V_1V_2 \leq 0$. It can be easily seen (or, see [6]) that the supremum of the risk of ψ is $\frac{1}{2}$. Note that there are many such orthogonal transformations L which will

satisfy the desired property for (EV_1, EV_2). It may be shown that neither of the rules φ_L^0 and ψ dominates the other. However, the characterization of the class of all admissible minimax rules is not known.

Now, instead of the zero-one loss function consider a loss function which takes the value 0 for correct decisions and equals $\ell(|\mu_1-\mu_2|)$ for any incorrect decision, where ℓ is a positive-valued bounded, continuous function such that $\ell(\Delta) \to 0$ as $\Delta \downarrow 0$. DasGupta and Bhattacharya [3] have shown that φ_L^0 is the unique minimax rule (and Bayes admissible) for the above loss function when $n_1 = n_2$.

It is clear that neither of φ_p^0 and φ_L^0 dominates the other. It is believed that φ_p^0 is also admissible.

2.1.2 Invariant Rules. Let us now consider the following conditions on the rules based on U_1, U_2, U_3:

Translation invariance:

$$\varphi(u_1, u_2, u_3) = \varphi(u_1, u_2, u_3 + b) \tag{2.11}$$

for all u_1, u_2, u_3 and $b \in R$.

A set of maximal invariants for (2.11) is given by (U_1, U_2). Hence we shall write a translation-invariant rule as a function of U_1 and U_2.

Sign invariance:

$$\varphi(u_1, u_2, u_3) = \varphi(-u_1, -u_2, -u_3) \tag{2.12}$$

for all u_1, u_2, u_3.

A translation-invariant rule is sign-invariant iff it is a function of $(u_1u_2/|u_2| \ , \ |u_2|)$. [10].

Symmetry:

$$\varphi_1(u_1,-u_2,u_3) = \varphi_2(u_1,u_2,u_3) \tag{2.13}$$

for all u_1, u_2 and u_3.

It is clear that both Ω_1 and Ω_2 are unchanged under the transformations $(u_1, u_2, u_3) \to (u_1, u_2, u_3 + c)$ and $(u_1, u_2, u_3) \to (-u_1, -u_2, -u_3)$. In terms of x, $\bar{x}_1$ and $\bar{x}_2$ these transformations are respectively $(x, \bar{x}_1, \bar{x}_2) \to (x+b, \bar{x}_1+b, \bar{x}_2+b)$ and $(x, \bar{x}_1, \bar{x}_2) \to (-x, -\bar{x}_1, -\bar{x}_2)$. The sets Ω_1 and Ω_2 are interchanged under the transformation $(u_1, u_2, u_3) \to (u_1, -u_2, u_3)$. This transformation is obtained by interchanging $(\bar{x}_1, n_1)$ and $(\bar{x}_2, n_2)$. We shall now show that φ_L^0 is the uniformly best translation-invariant, sign-invariant symmetric rule. For $(\nu_1,\nu_2,\nu_3) \in \Omega_1$ (i.e., $\nu_1 = \nu$, $\nu_2 = -c\nu$) the risk of a translation-invariant, sign-invariant symmetric rule φ is given by

$$
\begin{aligned}
&E_{\nu_1,\nu_2,\nu_3}\, \varphi_2(U_1, U_2) \\
&\quad = \int_0^\infty \int_0^\infty [\varphi_2(u_1,u_2)n(u_1;\nu)n(u_2;-c\nu) \\
&\qquad + \varphi_2(u_1,u_2)n(u_1;-\nu)n(u_2,c\nu) \\
&\qquad + \{1 - \varphi_2(u_1,u_2)\}n(u_1;-\nu)n(u_2;-c\nu) \\
&\qquad + \{1 - \varphi_2(u_1,u_2)\}n(u_1;\nu)n(u_2;c\nu)] \cdot du_1 du_2, \qquad (2.14)
\end{aligned}
$$

where $n(u;\nu)$ is the density of $N(\nu,1)$ at u. It may be seen that (2.14) is minimum (uniformly in ν and ν_3) for $\varphi_2(u_1,u_2) = 1$ when $u_1u_2 > 0$. The above result can also be proved using the distribution of $(U_1U_2/|U_2|,|U_2|)$ [10].

Kinderman [10] characterized the (essential) complete class among all translation-invariant, sign-invariant rules when $n_1 = n_2$.

2.1.3 Best Invariant Similar Test. The classification problem may be viewed in the light of Neyman-Pearson Theory. We may pose the problem as testing the hypothesis $H_1: \theta \in \Theta_1$ against the alternative $\theta \in \Theta_2$. We restrict our attention to the class of tests which are translation-invariant and

sign-invariant. Let ψ be a test function, i.e. $\psi(X, \bar{X}_1, \bar{X}_2)$ is the probability of rejecting H_1 given X, $\bar{X}_1$ and $\bar{X}_2$. Define

$$Y_1 = (1+1/n_1)^{-\frac{1}{2}}(X-\bar{X}_1), \tag{2.15}$$

$$Y_2 = [(1+1/n_2)^{-\frac{1}{2}}(X-\bar{X}_2) - (1+1/n_2)^{-\frac{1}{2}}(1+1/n_1)^{-1}(X-\bar{X}_1)]d, \tag{2.16}$$

$$Y_3 = (1+n_1+n_2)^{-\frac{1}{2}}(X+n_1\bar{X}_1+n_2\bar{X}_2), \tag{2.17}$$

where d is a constant chosen appropriately to make $\mathrm{Var}(Y_2) = 1$. If ψ is translation-invariant it will depend only on Y_1 and Y_2. Furthermore the sign-invariance of ψ means

$$\psi(y_1,y_2) = \psi(-y_1,-y_2). \tag{2.18}$$

Under H_1 the means of Y_1 and Y_2 are given by

$$\delta_1 \equiv EY_1 = 0, \quad \delta_2 = EY_2 = d\,(1+1/n_2)^{-\frac{1}{2}}(\mu_1-\mu_2). \tag{2.19}$$

Similarly, the means of Y_1 and Y_2 under H_2 are given by

$$\delta_1 = (1+1/n_1)^{-\frac{1}{2}}(\mu_2-\mu_1), \quad \delta_2 = -d(1+1/n_2)^{-\frac{1}{2}}(1+1/n_1)^{-1}(\mu_2-\mu_1). \tag{2.20}$$

In terms of δ_1 and δ_2 the parameter sets may be expressed as

$$\Delta_1 = \{(\delta_1,\delta_2): \delta_1 = 0, \delta_2 \neq 0\}, \tag{2.21}$$

$$\Delta_2 = \{(\delta_1,\delta_2): \delta_2 = a\delta_1 \neq 0\} \tag{2.22}$$

under H_1 and H_2, respectively; $a = -d(1+1/n_1)^{-\frac{1}{2}}(1+1/n_2)^{-\frac{1}{2}}$. Since δ_2 is still unknown under H_1 we require ψ to be similar size α for H_1, i.e.

$$E_{0,\delta_2}\psi(Y_1,Y_2) = \alpha \quad \text{for all} \quad \delta_2 \neq 0. \tag{2.23}$$

This is equivalent to

$$\int_{-\infty}^{\infty} \psi(y_1,y_2)n(y_1;0)dy_1 = \alpha \qquad \text{a.e. } (y_2). \tag{2.24}$$

The power of the test ψ is given by

$$E_{\delta_1,\delta_2}\psi(Y_1,Y_2) = \frac{1}{2}[E_{\delta_1,\delta_2}\psi(Y_1,Y_2) + E_{-\delta_1,-\delta_2}\psi(Y_1,Y_2)]$$

$$= \frac{1}{2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-\delta_1^2(1+a^2)/2} n(y_1;0)n(y_2;0)\psi(y_1,y_2)$$

$$\left[e^{\delta_1(y_1+ay_2)} + e^{-\delta_1(y_1+ay_2)}\right]dy_1dy_2 \quad . \tag{2.25}$$

Using the Neyman-Pearson Lemma in order to maximize

$$\int_{-N}^{\infty} \psi(y_1,y_2)\left[e^{\delta_1(y_1+ay_2)} + e^{-\delta_1(y_1+ay_2)}\right]n(y_1,0)dy_1 \quad . \tag{2.26}$$

subject to (2.24) we get the following optimum test:

$$\psi^*(y_1,y_2) = 1 \quad \text{iff} \quad |y_1+ay_2| > k(y_2), \tag{2.27}$$

where $k(y_2)$ is chosen so that

$$\int_{-ay_2-k(y_2)}^{-ay_2+k(y_2)} n(y_1;0)dy_1 = 1-\alpha \quad . \tag{2.28}$$

Thus ψ^* is the uniformly most powerful invariant similar test. The above result is due to Schaafsma [12].

2.2 The common variance σ^2 is unknown

It may be easily seen that the rules given by (2.7) and (2.8) are still unique Bayes. Moreover, the rule ψ_L^1 is the one which accepts $\theta \in \Theta_1$ if (2.10) holds and it is admissible minimax. When $n_1 = n_2$ Das Gupta and Bhattacharya [3] have shown that the rule ψ_L^1 is the unique (a.e.) minimax when the loss for incorrect decision is $\ell(|\mu_1-\mu_2|/\sigma)$, where ℓ is a positive valued, bounded, continuous function such that $\ell(\Delta) \to 0$ as $\Delta \downarrow 0$. To see all the above results, not that (U_1,U_2,U_3,S) are sufficient statistics in this case and S is distributed independently of (U_1,U_2,U_3). It also follows that ψ_L^I is the uniformly best translation-invariant, symmetric rule. To see this, condition on S and fix σ.

Schaafsma [13] has shown that the following critical region for testing H_1 against H_2 is (i) similar of size α for H_1, (ii) unbiased for H_2, and

(iii) asymptotically (as $\min(n_1,n_2) \to \infty$) most stringent among all level α tests:

$$(2.29) \qquad Y_1 \operatorname{sign}(Y_2) > \sqrt{S}\, t_{n_1+n_2-2,\alpha} ,$$

where Y_1 and Y_2 are given in (2.15) and (2.16), S is given in (1.5), and $t_{n_1+n_2-2,\alpha}$ is the upper $100\alpha\%$ point of the Student's t distribution with n_1+n_2-2 degrees of freedom. However, it is very likely that this test is not admissible.

It follows from Kiefer and Schwartz [9] that the rule ψ_L^λ is a (unique) Bayes rule. We shall give a sketch of the prior distribution against which ψ_L^λ is unique Bayes. Consider U_1,U_2,U_3 as defined in (2.1) - (2.3). Then U_i's are independently distributed, and $U_1 \sim N(\nu_i, \sigma^2)$. Moreover, under $\theta \in \Theta_i$ (i.e. $(\nu_1,\nu_2,\nu_3) \in \Omega_i$) we have $\nu_1 = \nu$, $\nu_2 = (-1)^i c\nu$, $\nu \neq 0$. The prior distribution is given as follows:

(i) $P(\theta \in \Theta_i) = \xi_i$, $i = 1,2$.

(ii) Given $\theta \in \Theta_i$, the conditional distribution of (ν,ν_3,σ^2) is derived from the following:

(iia) Given $\sigma^2 = (1+\tau^2)^{-1}$, the conditional distribution of $(\frac{\nu}{\sigma^2}, \frac{\nu_3}{\sigma^2})$ is the same as that of $(\tau V, \tau V_3)$, where V and V_3 are independently distributed with $V \sim N(0, (1+\tau^2)/(1+c^2))$ and $V_3 \sim N(0, 1+\tau^2)$.

(iib) The density of τ is proportional to $(1+\tau^2)^{-(m+1)/2}$.

3. Multivariate Case: Σ known

Without any loss of generality we shall assume that $\Sigma = I_p$. First we shall derive a class of Bayes rules and obtain an admissible minimax rule. Define U_1, U_2, U_3 and k_1, k_2 as in (2.]) - (2.3), except that U_i's are now $p \times 1$ vectors and $U_i \sim N_p(\nu, I_p)$. Correspondingly redefine the sets Ω_i as follows:

$$(3.1) \qquad \Omega_i = \{(\nu_1, \nu_2, \nu_3): \ \nu_1 = \nu, \nu_2 = (-1)^i c\nu, \neq 0; \ \nu, \nu_3 \in R^p\},$$

$i = 1,2$. As before U_3 may be eliminated from a Bayes rule by taking a fixed distribution, independent of (ν_1, ν_2), under both Ω_1 and Ω_2. Now consider the prior distribution which assigns the probability ξ_i to Ω_i and, given $\nu_1 = \nu, \nu_2 = (-1)^i c\nu$, the distribution of ν is $N_p(0, \tau^2 I_p)$. It can now be seen that the unique (a.e.) Bayes rule against the above prior distribution decides $(\nu_1, \nu_2, \nu_3) \in \Omega_1$ iff

$$(3.2) \qquad U_1' U_2 \leq k,$$

where k is a function of ξ_1 and ξ_2; conversely, given k the probability ξ_1 and ξ_2 can be suitably chosen. Thus any likelihood-ratio rule φ_L^k is Bayes and admissible.

We shall now show that φ_L^0 is minimax. First we shall consider a different prior distribution against which φ_L^0 is unique Bayes. As before, ν_3 can be eliminated from the problem. Now consider a prior distribution which assigns equal probabilities to the sets Ω_1^* and Ω_2^*, where

$$(3.3) \qquad \Omega^* = \{(\nu_1, \nu_2): \ \nu_1 = \nu, \nu_2 = (-1)^i c\nu, \ \nu \neq 0, \ \nu \in R^p\}.$$

Moreover, given that $(\nu_1,\nu_2) \in \Omega_i^*$, the distribution of ν is taken to be uniform over the surface of the hypershpere $\nu' \nu = \Delta^2$. See Das Gupta [4] to get a detailed proof of the fact that ϕ_L^0 is unique (a.e.) Bayes against the above prior distribution. To see that ϕ_L^0 is minimax, note that the risk of ϕ_L^0 is constant over the set

$$(3.4) \qquad \{(\nu_1,\nu_2,\nu_3): \ \nu_1 = \nu, \nu_2 = -c\nu, \ \nu' \nu = \Delta^2\}.$$
$$\cup \ \{(\nu_1,\nu_2,\nu_3): \ \nu_1 = \nu, \nu_2 = c\nu, \ \nu' \nu = \Delta^2\}.$$

Das Gupta [4] has also shown that the rule ϕ_L^0 is the unique (a.e.) minimax when the loss for any correct decision is zero, and the loss for deciding $\mu = \mu_i$ incorrectly is

$$(3.5) \qquad \ell[(1 + 1/n_i)^{-1} (\mu - \mu_i)' (\mu - \mu_i)],$$

where ℓ is a positive-valued, bounded, continuous function such that $\ell(\Delta) \to 0$ as $\Delta \downarrow 0$.

As in (2.11) we may call a rule ϕ translation-invariant if

$$(3.6) \qquad \phi(U_1,U_2,U_3) = \phi(U_1,U_2,U_3 + b),$$

for all $b \in R^p$. Clearly, (U_1,U_2) is a set of maximal invariants. A rule ϕ is called orthogonally-invariant if

$$(3.8) \qquad \phi(U_1,U_2,U_3) = \phi(OU_1,OU_2,OU_3),$$

for all orthogonal $p \times p$ matrices O.

Kudo [11] considered the following "symmetry" condition for a translation-invariant rule ϕ:

$$(3.9) \qquad \beta_1(\phi;(1 + 1/n_2)^{-1/2}d) = \beta_2(\phi;(1 + 1/n_1)^{-1/2}d),$$

where $\beta_i(\phi;d) = E_\theta \phi_i$ when $d = (\mu_1 - \mu_2)$ and $\mu = \mu_i$. Moreover, he

required $\beta_i(\phi;d)$ to depend on d only through $d'd$. This condition clearly holds if ϕ is translation-invariant and orthogonally-invariant. Note also that for a translation-invariant and an orthogonally-invariant rule ϕ satisfying (2.13) the condition (3.9) holds. Kudo [11] has shown that ϕ_L^0 simultaneously maximizes both $\beta_1(\phi;d)$ and $\beta_2(\phi;d)$ in the class of all translation-invariant rules satisfying (3.9) and for which $\beta_i(\phi;d)$ depends on d only through $d'd$. This can be seen easily by integrating the probability of correct classification with respect to the uniform distribution of ν over $\nu'\nu = \Delta^2$, where $\nu_1 = \nu, \nu_2 = (-1)^i c\nu$.

Rao [15] has considered the class $\underline{\Phi}^*$ of rules whose probabilities of misclassification depend only on

$$\Delta^2 = (\mu_1 - \mu_2)' \Sigma^{-1} (\mu_1 - \mu_2). \tag{3.10}$$

For a rule $\phi \in \underline{\Phi}^*$ let $G_1(\phi;\Delta^2)$ and $G_2(\phi;\Delta^2)$ be the error probabilities when $\mu = \mu_1$ and $\mu = \mu_2$, respectively. Rao [15] has posed the problem of minimizing

$$\frac{d}{d\Delta^2} \{aG_1(\phi;\Delta^2) + bG_2(\phi;\Delta^2)\}\Big|_{\Delta = 0}, \tag{3.11}$$

subject to the condition that the ratio of $G_1(\phi;0)$ to $G_2(\phi;0)$ is equal to some specified constant. The resulting optimum rule decides $\mu = \mu_1$ iff

$$\begin{aligned} a[(X-\bar{X}_1) - (1+1/n_1)(X-\bar{X}_2)]' [(X-\bar{X}_1) - (1+1/n_1)(X-\bar{X}_2)] \\ - b[(1+1/n_2)(X-\bar{X}_1) - (X-\bar{X}_2)]' [(1+1/n_2)(X-\bar{X}_1) - (X-\bar{X}_2)] \geq k \end{aligned} \tag{3.12}$$

The above rule coincides with ϕ_L^0 when $n_1 = n_2$ and $a = b$, $k = 0$.

4. Multivariate Case: Σ unknown

First we shall show that a likelihood-ratio rule ψ_L^{λ} is unique (a.e.) Bayes and hence is admissible (for zero-one loss function). Note that U_1, U_2, U_3 and S are sufficient statistics in this case, where U_i's (in $p \times 1$ vector notations) are given by (2.1) - (2.3) and S is given by (1.5). Here $U_i \sim N_p(\nu_i, \Sigma)$. We now consider the following prior distribution.

(i) $\underline{P}\ (\theta \in \Theta_i) = \xi_i$

(ii) Given $\theta \in \Theta_i$ (i.e., $\nu_1 = \nu, \nu_2 = (-1)^i c\nu$), the conditional distribution of (ν, ν_3, Σ) is derived from the following:

(iia) Given $\Sigma^{-1} = I_p + \tau\tau'(\tau:\ p \times 1)$, the conditional distribution of $(\Sigma^{-1}\nu,\ \Sigma^{-1}\nu_3)$ is the same as the distribution of $(\tau V,\ \tau V_3)$, where V and V_3 are independently distributed as

$$N(0,\ (1+c^2)^{-1}\ (1+\tau'\tau)) \text{ and } N(0,\ 1+\tau'\tau),$$

respectively.

(iib) The density of τ is proportional to $(1+\tau'\tau)^{-(m+1)/2}$, where $m > p - 1$.

Following a simplified version of the results of Kiefer and Schwartz [9] it can be shown that a unique (a.e.) Bayes rule against the above prior distribution accepts $\mu = \mu_1$ if (1.7) holds, where λ is a function of ξ_i's; conversely, given λ the constants ξ_i's can by appropriately chosen.

Das Gupta [4] has considered a class $\underline{\phi}^{**}$ of rules invariant under the following transformations:

$$(4.1 \qquad (X, \bar{X}_1, \bar{X}_2, S) \to (AX + b, A\bar{X}_1 + b,\ A\bar{X}_2 + b,\ ASA'),$$

where A is any $p \times p$ nonsingular matrix and b is any vector in R^p. It is shown [4] that a set of maximal invariants is given by (m_{11}, m_{12}, m_{22}), where

$$m_{ij} = U_i' S^{-1} U_j / m. \tag{4.2}$$

When $\nu_1 = \nu, \nu_2 = (-1)^i c\nu$, $\nu' \Sigma^{-1} \nu = \Delta^2$, the joint density of (m_{11}, m_{12}, m_{22}) is given by [14]

$$p_i(m_{11}, m_{12}, m_{22}; \Delta^2) = K \exp[-\Delta^2(1+c^2)/2] \, |M|^{(p-3)/2} \sum_{j=0}^{\infty} g_j (\tfrac{1}{2}\Delta^2)^j \, h_j(m_{11}, m_{12}, m_{22}), \tag{4.3}$$

where

$$h_j(m_{11}, m_{12}, m_{22}) = \frac{(m_{11} + 2(-1)^i c m_{12} + c^2 m_{22} + (1 + c^2) \, |M|)^j}{|I_2 + M|^{\frac{1}{2}(m+2)+j}} \tag{4.4}$$

$$|M| = \det M, \quad M = \begin{pmatrix} m_{11} & m_{12} \\ m_{12} & m_{22} \end{pmatrix}, \tag{4.5}$$

$$m = n_1 + n_2 - 2, \tag{4.6}$$

and $K > 0$, $g_i > 0$ are numerical constants.

Consider a prior distribution which assign equal probabilities to Θ_i and, given $\theta \in \Theta_i$ (i.e. $\nu_1 = \nu, \nu_2 = (-1)^i c\nu$) the value of $\nu' \Sigma^{-1} \nu = \Delta^2$ is held fixed. The Bayes rule in $\underline{\Phi}^{**}$ against the above prior distribution decides $\theta \in \Theta_1$ iff

$$m_{12} < 0 \tag{4.7}$$

To see this, note that for $a > 0$

$$(a + x)^j < (a - x)^j \tag{4.8}$$

for any positive j if $x < 0$. The relation (4.7) is the same as (1.7) for $\lambda = 1$. It now follows easily that the rule ψ_L^1 is admissible and minimax in $\underline{\phi}^{**}$ [4]. Das Gupta [4] has also shown that ψ_L^1 is the unique (a.e.) minimax in $\underline{\phi}^{**}$ if the loss for any correct decision is zero and the loss for deciding $\mu = \mu_i$ incorrectly is

$$(4.9) \qquad \ell[(1 + 1/n_i)^{-1} (\mu - \mu_i)' \Sigma^{-1} (\mu - \mu_i)],$$

where ℓ is a positive-valued, bounded, continuous function such that $\ell(\Delta) \to 0$ as $\Delta \downarrow 0$.

Again for this case Rao [15] considered the class $\underline{\phi}^{**}$ of rules whose probabilities of misclassification depend only on Δ^2 given in (3.10). Then he derived the optimum rule which minimizes the expression given by (3.11) subject to the condition of similarity for the subset of the parameters given by $\mu_1 = \mu_2$. The optimum rule decides $\mu = \mu_1$ iff

$$(4.10) \qquad a[(X-\bar{X}_1) - (1 + 1/n_1)(X-\bar{X}_2)]' B^{-1}[(X-\bar{X}_1) - (1 + 1/n_1)(X-\bar{X}_2)]$$
$$- b[(X-\bar{X}_2) - (1 + 1/n_2)(X-\bar{X}_1)]' B^{-1}[(X-\bar{X}_2) - (1 + 1/n_2)(X-\bar{X}_1)]$$
$$\geq c(B),$$

where

$$(4.11) \qquad B = mS + \frac{n_1 n_2}{1+n_1+n_2} [(1 + 1/n_2)(X-\bar{X}_1)'(X-\bar{X}_1)$$
$$+ (1 + 1/n_1)(X-\bar{X}_2)'(X-\bar{X}_2) - 2(X-\bar{X}_1)'(X-\bar{X}_2)].$$

It is not clear why Rao imposed the similarity condition even after restricting to the class $\underline{\phi}^{*}$. One may directly consider the class of rules invariant under (4.1) and try to minimize (3.11) subject to the condition that $G_i(\phi;0)$ is equal to a specified constant. Using (4.3) it can be found that the optimum rule decides μ μ_1 iff

$$(4.12)\quad a(k_1^2 m_{11} + k_2^2 m_{22} + (k_1^2 + k_2^2)\,|M| - 2k_1k_2m_{12})(1 + 1/n_2)^{-1}$$

$$- b(k_1^2 m_{11} + k_2^2 m_{22} + (k_1^2 + k_2^2)\,|M| + 2k_1k_2m_{12})(1 + 1/n_1)^{-1}$$

$$> \lambda \det (I_2 + M).$$

As in (2.29) a similar region for Θ_1 may be constructed for this case also. It is given by the following:

$$(4.13)\quad Y_2' (mS + Y_1Y_1')^{-1} Y_1/[Y_2' (mS + Y_1Y_1')^{-1} Y_2]^{\frac{1}{2}} > k,$$

where Y_1 and Y_2 are given in (2.15) and (2.16) in vector notations.

5. Multivariate Case: μ_1 and μ_2 known

In this case the plug-in rules are given by the following: Decide $\mu = \mu_1$ if

$$(4.14) \qquad (X - \mu_1)' A^{-1}(X - \mu_1) - (X - \mu_2)' A^{-1}(X - \mu_2) > \lambda,$$

where

$$(4.15) \qquad A = [mS + n_1(\bar{X}_1 - \mu_1)(\bar{X}_1 - \mu_1)' + n_2(\bar{X}_2 - \mu_2)(\bar{X}_2 - \mu_2)].$$

On the other hand, a likelihood-ratio rule decides $\mu = \mu_1$ iff

$$(4.16) \qquad \frac{1 + (X - \mu_2)' A^{-1}(X - \mu_2)}{1 + (X - \mu_1)' A^{-1}(X - \mu_1)} > \lambda \ (0 < \lambda).$$

Define $m^* = m + 2$.

Without loss of generality we may assume that $\mu_1 = 0$ and $\mu_2' = (1,0,\ldots 0)$. Then the problem is invariant under the following transformations:

$$(4.17) \qquad (X,A) \to (L X, L A L'),$$

where L is a nonsingular $p \times p$ matrix of the form

$$(4.18) \qquad L = \left[\begin{array}{c|c} 1 & L_{12} \\ \hline 0 & L_{22} \end{array}\right]$$

It can be seen that a set of maximal invariants is given by $(X_{1.2}, X'_{(2)}A_{22}^{-1}X_{(2)}, A_{11.2})$, where

$$(4.19) \qquad A = \left[\begin{array}{c|c} A_{11} & A_{12} \\ \hline A_{21} & A_{22} \end{array}\right] \begin{matrix} 1 \\ p-1 \end{matrix}, \quad X = \begin{pmatrix} X_1 \\ X_{(2)} \end{pmatrix} \begin{matrix} 1 \\ p-1 \end{matrix}$$

$$\begin{matrix} 1 & p-1 \end{matrix}$$

$$(4.20) \qquad A_{11.2} = A_{11} - A_{12}\, A_{22}^{-1}\, A_{21},$$

$$(4.21) \qquad X_{1.2} = X_1 - A_{12}\, A_{22}^{-1}\, X_{(2)}$$

$A_{11.2}$ is distributed, independently of $(X_{1.2}, X'_{(2)}A_{22}^{-1}X_{12})$, as $\sigma_{11.2}\chi^2_{m^*-p+1}$; given $X'_{(2)}A_{22}^{-1}X_{(2)}$, the distribution of $X_{1.2}$ is $N(d, \sigma_{11.2}(1 + X'_{(2)}A_{22}^{-1}X_{(2)}))$, and $X'_{(2)}A_{22}^{-1}X_{(2)}$ is distributed as the ratio of independent χ^2_{p-1} and $\chi^2_{m^*-p+2}$ variates. In the above d is equal to 0 or 1 according as $\mu = \mu_1$ or $\mu = \mu_2$, and $\sigma_{11.2}$ is the residual variance of X_1 given $X_{(2)}$. It can be shown now that the following rule is minimax (and Bayes) in the class of rules invariant under (4.18): Decide $\mu = \mu_1$ iff

$$(4.22) \qquad X_{1.2} < 1/2.$$

The relation (4.22) is the same as (4.14) for $\lambda = 0$, and as (4.16) for $\lambda = 1$. The above region is not similar for $\mu = \mu_1$. Such a simular region may be constructed using

$$(4.23) \qquad X_{1.2}(1 + X'_{(2)}A_{22}^{-1}X_{(2)})^{-\frac{1}{2}} (A_{11.2}/(m^*-p+1))^{-\frac{1}{2}}$$

which is distributed as Student's' t - distribution with m^*-p+1 degrees of freedom when $\mu = \mu_1$. The Mahalanobis distance is equal to $(\sigma_{11.2})^{-\frac{1}{2}}$

in this case. The probabilities of correct classification for the rule given by (4.22) are the same and they decrease as p increases if $\sigma_{11.2}$ is held fixed.

This section is new in the literature and it is due to the present author.

Referneces

[1] Anderson, T.W. (1951). Classification by multivariate analysis. Psychometrika 16, 31-50.

[2] Anderson, T.W. (1958). An Introduction to Multivariate Statistical Analysis. Wiley, New York.

[3] Das Gupta S. and Bhattacharya, P.K. (1964). Classification into exponential populations. Sankhya, Ser. A, 26, 17-24.

[4] Das Gupta, S. (1965). Optimum classification rules for classification into two multivariate normal populations. Ann. Math. Statist. 36, 1174-1184.

[5] Das Gupta S. and Kinderman A. (1974). Classifiability and designs for sampling. Sankhya Ser. A, 36, 237-250.

[6] Das Gupta, S. (1974). Probability inequalities and errors in classification. Ann. Statist., 2, 751-762.

[7] Fisher, R.A. (1936). The use of multiple measurements in taxonomic problems. Ann. Eugenics 7, 179-188.

[8] Fisher, R.A. (1938). The statistical utilization of multiple measurements. Ann. Eugenics 8, 376-386.

[9] Kiefer, J. and Schwartz, R. (1956). Admissible Bayes character of T^2-, R^2-, and other fully invariant tests for classical multivariate normal problems. Ann. Math. Statist. 36, 747-770.

[10] Kinderman, A. (1972). On some problems in classification. Tech. Rep. 178, School of Statistics, University of Minnesota, Minneapolis.

[11] Kudo, A. (1959). The classificatory problem viewed as a two-decision problem. Memoirs of Faculty of Science Kuyushu University, Ser. A, 13, 96-125.

[12] Schaafsma, W. (1971). Testing statistical hypotheses concerning the expectations of two independent normals; both with variance 1. I and II. Proc. Kon. Ned. Ak. (Indag. Math.) 33, 86-105.

[13] Schaafsma, W. and Van Verk, G.N. (1977). Classification and discrimination problems with applications. Statistica Neerlandica, 31, 25-45.

[14] Sitgreaves, R. (1952). On the distribution of two random matrices used in classification procedures. Ann. Math. Statist. 23, 263-270.

[15] Rao, C.R. (1954). A general theory of discrimination when the information about alternative population distributions is based on samples. Ann. Math. Statist., 25, 651-670.

[16] Von Mises, R. (1945). On the classification of observation data into distinct groups. Ann. Math. Statist., 16, 68-73.

[17] Wald, A. (1944). On a statistical problem arising in the classification of an individual into one of two groups. Ann. Math. Statist., 15, 145-162.